Översättningssystemet T4F
- en implementation för ATIS-domänen

Maria Holmqvist
x02marho@ida.liu.se
Linköpings universitet, IDA
24 april 2003

1 Introduktion

T4F står för tokenisering, taggning, transfer, transformering och filtrering, och är ett system för korpusbaserad översättning från engelska till svenska.

1.1 Superlänkar

Översättningen är lexikalistisk och bygger på att varje ord förses med s.k. supertaggar. En supertagg innehåller både lingvistisk/semantisk information om ordet i sig samt en beskrivning av ordets lokala kontext eftersom en korrekt översättning för det mesta beror på olika faktorer i ordens omgivning. De möjliga översättningarna kommer att begränsas då inte bara källordet ska korrespondera med målordet utan även käll- och målkontext ska korrespondera med varandra.

En ramverk för översättning baserad på superlänkar presenteras i Ahrenberg (2000). En superlänk definieras som ett par av supertaggar där den första taggen innehåller kontexten för källordet och den andra kontexten för målordet. I detta ramverk behöver inte supertaggarna bestå av kompletta beskrivningar av kontexten. En supertagg kan bestå av flera olika slags information, egentligen allt som kan vara avgörande för hur ett ord översätts. Dependensrelationer, spatiala relationer, semantiska bestämningar eller explicita operationer är möjliga data i en supertagg.

1.2 SCOTS

Den första realiseringen av superlänkbaserad översättning är systemet SCOTS – Superlink COnstrained Transfer lexicalist translation System (Jonsson, 2001). I SCOTS utgörs supertaggarna i grunden av trigramtaggar som består av ordets egen tagg (ordklass, särdrag och semantik) samt taggen för ordet till höger och taggen för ordet till vänster. Supertaggarna kan även innehålla information om andra spatiala relationer vilka läggs till manuellt i lexikon.

I T4F utnyttjas dependensrelationer av olika slag utöver spatiala relationer för att definiera supertaggar. Dependensträdet kommer från taggning med FDG – Functional Dependency Grammar från Connexor (Tapanainen och Järvinen, 1997). En annan skillnad mot SCOTS är att transformeringssteget i T4F sker utifrån explicita ordningsregler.

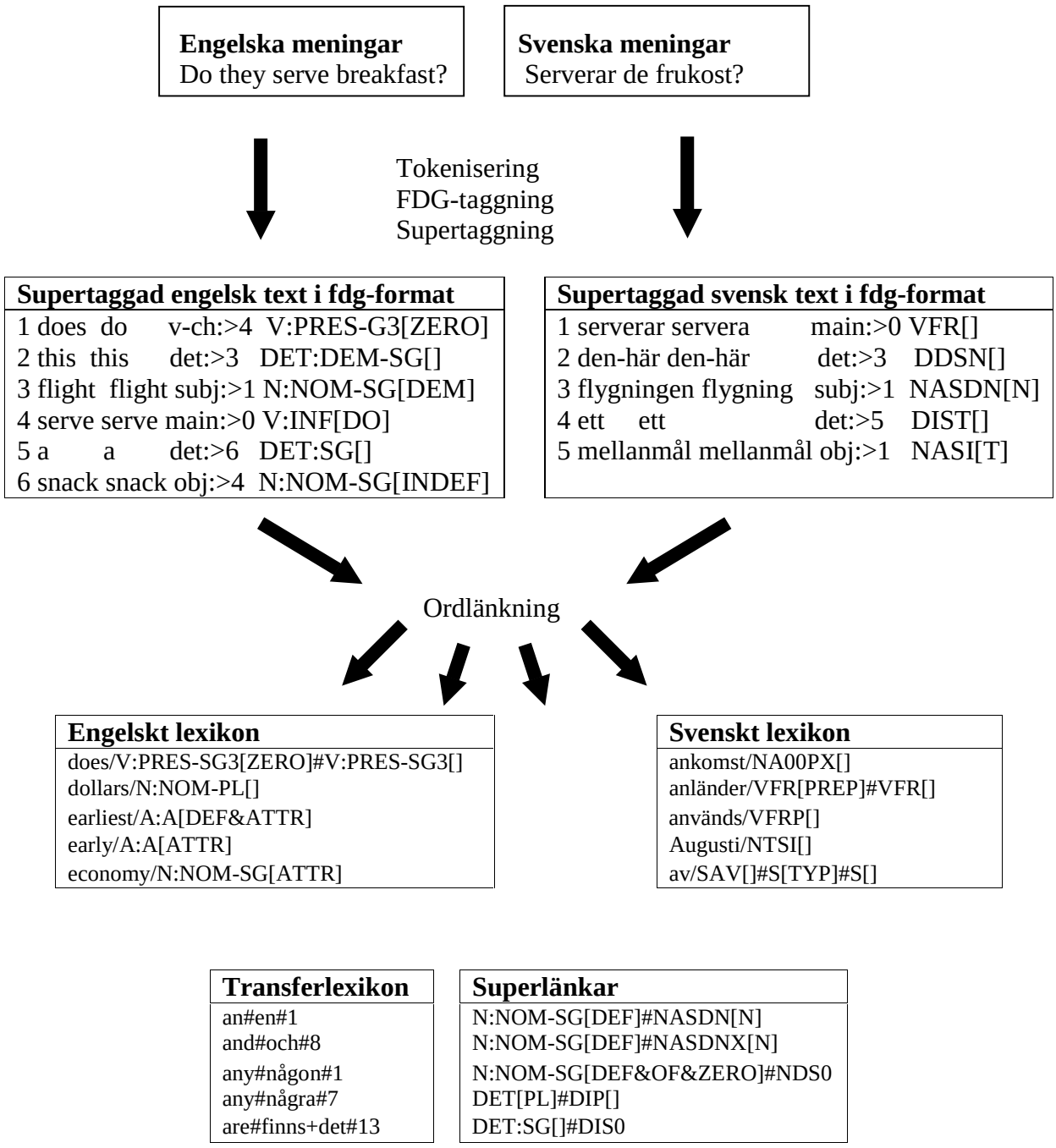
För att jämföra dessa två metoder utvecklades översättningssystemet (T4F) utifrån förutsättningar som så långt som möjligt liknade SCOTS. Samma texter och en likadan ordlänkning användes. Målet var att de ursprungliga taggarna (innan supertaggningen) skulle ha samma granularitet som ATIS-taggarna i SCOTS.

1.3 Texter

Systemet är uppbyggt från meningslänkade ATIS-texter på engelska och svenska översatta med det automatiska översättningssystemet SLT: *S*poken *L*anguage *T*ranslator, samt förbättrade för att få acceptabla svenska översättningar (Jonsson, 2001). Texten delades in i en träningsdel och en testdel identiska med de som användes i utveckling och testning av SCOTS.

1.4 Översättning

De lexikon som behövs för översättning utvinns ur träningsdata. Nedan följer en schematisk beskrivning av hur detta går till i T4F.



2 Resurser för översättningssystemet

Resursfilerna till översättningssystemet består av ett transferlexikon, ett lexikon med superlänkar samt två enspråkiga lexikon, ett för källspråket och ett för målspråket. För att skapa dessa resursfiler behövs två grundfiler, en engelsk och en svensk, bestående av FDG-taggade meningar med supertaggar tillagda. Dessa filer skapas utifrån två meningslänkade textfiler så här (indata och utdata är markerad i fetstil):

Engelska		**Svenska**	
Steg	Verktyg och resurser	Steg	Verktyg och resurser
Tokenisering	Tokeniseringsscript **eng-trn.txt**	Tokenisering	Tokeniserigsscript MWU-fil **swe-trn.txt**
FDG-taggning	Taggningsscript. FDG Anpassningar till FDGs lexikon	FDG-taggning	Taggningsscript. FDG
Ommappning av taggar.	Taggningsscriptet	Omtaggning	ATIS-taggad fil taggad med FDG Mappningsscript
Supertaggning	Taggningsscript Supertaggningsregler Semantiska kategorier	Supertaggning	Taggningsscript Supertaggningsregler för svenska Semantiska kategorier för svenska
Resultat	**eng-trn-supertagged.fdg**	Resultat	**swe-trn-supertagged.fdg**

Figur 1. Förbehandling och supertaggning.

2.1 Tokenisering

Engelska och svenska träningsmeningar tokeniserades så att punkter och andra skiljetecken togs bort. I den svenska delen delades sammansättningar som ”Atlanta-flygplatsen” upp. Sedan sammanfördes flerordsenheter med ett bindestreck för att behandlas som ett token av FDG. I den engelska delen fördes endast ett fåtal flerordsenheter ihop i detta steg. Resten togs istället hand om av FDG med ett anpassat lexikon.

2.2 Anpassning av engelskt lexikon i FDG

I den engelska versionen av FDG kan egna ord läggas till i ett separat lexikon som läses in och används av FDG under parsning. Detta utnyttjades bl.a. för att lägga in semantiska taggar. Här lades sådant som Atlanta - Location och Delta Airlines - Company in. Genom det anpassade lexikonet kan man också sammanföra MWU'er t.ex.

first_class N
one_way N

samt undertrycka möjliga tolkningar som förmodligen aldrig kommer förekomma inom domänen:

American -A
fare -V
fares -V
mean –N

De fördefinierade taggar som var möjliga att använda i det anpassade lexikonet var specificerade i dokumentationen till engelska FDG.

2.3 Taggning

Svensk och engelsk text taggades sedan med FDG vilket gav varje mening en grammatisk tolkning i form av ett dependensträd. Varje ord får dessutom en ordklasstagg med morfologiska särdrag. För att hålla systemet jämförbart med SCOTS mappades FDGs engelska ordklasstaggar om till kategorier som motsvarade de ATIS-taggar för engelska som användes i SCOTS.

För den svenska delen gjordes ingen ommappning av FDGs taggar, utan svenska ATIS-taggar överfördes direkt från den manuellt taggade svenska filen i SCOTS till den svenska FDG-filen ord för ord, vilket innebär att endast dependenserna behölls från den ursprungliga FDG-taggningen.

2.4 Supertaggning

Det avgörande steget i uppbyggnaden av systemet är supertaggningen, dvs. att med hjälp av bl.a. dependensträdet lägga till kontextuella särdrag till grundtaggningen. Ett nomen som hade en bestämd determinant (the) får t.ex. supertaggen [DEF]. Ordet ”the” i sin tur får supertaggen [PL] om den bestämmer ett nomen i plural eller [SG] om den bestämmer ett nomen i singular.

Figur 2. Dependensträd.

Efter länkning kommer transferlexikonet bl.a. innehålla ingångarna i figur 3.

```
flight/flygning#flygningen
the/den#det#de
```

Figur 3. Transferlexikon.

Superlänkarna i sin tur kommer se ut som i figur 4.

```
N:NOM-SG[DEF]#NADN[]
N:NOM-SG[DEF]#NADT[]
N:NOM-SG[]#NASI[]
DET:DEF-X[SG]#DDSN[]
DET:DEF-X[SG]#DDST[]
DET:DEF-X[PL]#DDP[]
```

Figur 4. Superlänkar.

Reglerna för supertaggningen har följande format. Varje regel berör endast en eller två noder i taget.

```
IF Lemma=the & Enode=Arg & Pos[e]=N:NOM-SG THEN (ADD c SG) & (ADD e DEF)
IF Lemma=the & Enode=Arg & Pos[e]=N:NOM-PL THEN (ADD c PL) & (ADD e DEF)
```

Figur 5. Exempel på supertaggningsregler.

Nästan alla kategorier från FDG kan användas som attribut i reglerna: graford, lemma, ordklass, funktionell relation, dependens/spatial relation. Även nuvarande supertaggar kan användas som attribut, vilket gör att ordningen på reglerna får betydelse.

Attribut	Betydelse	Ex. på värden (för engelska)
Form	Graford	the, flights, can, is, are
Lemma	Lemma	be, flight, price
Pos	Ordklass	A:A, N:NOM-SG, PREP:AV
Rel	Funktionell relation	tmp, subj, det, comp, obj, main
Cxt	Supertagg-attribut	DEF, LOC, ATTR
Enode	Anger relation mellan två noder	Arg, Ladj, L2adj
Sem	Semantisk kategori (från **seminfo**)	MONTH

Figur 6. Möjliga attribut i supertaggningsreglerna.

Attributet Enode anger vilken relation två noder har till varandra och kan anta följande värden:

Värde	Relation
Arg	Argument - Dependent till
Ladj	Left adjacent – Till vänster om
L2adj	Left 2 adjacent – Två ord till vänster om
Radj	Right adjacent – Till höger om
R2adj	Right 2 adjacent – Två ord till höger om

Figur 7. Värden till attributet Enode.

För ATIS-domänen behövdes ca 85 engelska supertaggningsregler.

2.5 Ordlänkning

För att skapa enspråkiga lexikon, transferlexikon och superlänkar användes *I*Link*, ett interaktivt ordlänkningsprogram som automatiskt skapar lexikonresurser utifrån sina länkfiler (se fig. 8) Länkfilerna skapades manuellt i *I*Link* och kan därför betraktas som perfekta.

2.5.1 Länkningsstrategi

Samma ordlänkningsstrategi, 1:N, användes som tidigare i SCOTS. Varje token på källspråkssidan måste länkas mot noll (en strykning) eller flera tokens på målspråkssidan. Däremot kunde inte flera engelska tokens länkas mot en eller flera enheter på svenska sidan. För att kunna länka en engelsk MWU till något på målspråkssidan var vi alltså tvungna att

sätta samman denna enhet redan i taggningssteget. Enheter på målspråkssidan kunde inte heller länkas som tillägg utan alla svenska ord måste hänga ihop med någon enhet på källsidan.

Två versioner av ordlänkning av träningskorpusen gjordes. En som så långt som möjligt liknade den tidigare i SCOTS (Alignment1) och en som var lite mer konsekvent men fortfarande följde samma principer (Alignment2).

	Resurser	Resultat
1	***I*Link*** Länkfil eng-trn-supertagged.fdg swe-trn-supertagged.fdg	Transferlexikon – transfer.bilex Superlänkar – superlinks.bilex
2	***I*Link*** (***Link reporter***) Länkfil eng-trn-supertagged.fdg swe-trn-supertagged.fdg	Engelsk lexikon – english.dict Svenskt lexikon – swedish.dict
3	*Efterbehandlingskript* transfer.bilex superlinks.bilex english.dict swedish.dict	transfer superlinks englex swelex

Figur 8. Steg vid skapande av resursfiler.

2.5.2 Transferlexikon och superlänkar

I *I*Link* skapas transferlexikon och superlänkar utifrån två supertaggade grundfiler och en länkfil så att varje källspråksenhet sattes samman med sin motsvarighet på målspråkssidan tillsammans med parets frekvens.

2.5.3 Enspråkiga lexikon

De enspråkiga resursfilerna skapas på liknande sätt genom att använda verktyget *LinkReporter* i *I*Link.* Med *LinkReporter* kan man välja ut de FDG-attribut man vill ha en lista över. För att skapa ett engelskt enspråkigt lexikon är det källspråkets graford och källspråkets ordklass- och supertaggar som ska listas. Genom att göra det här steget i *I*Link* kan man utnyttja länkfilen man gjort för att få flerordsenheter som en ingång i lexikon. Dessa får då en sammansatt tagg t.ex. finns+det/VFR[LOC]_PSST[]

Eftersom man inte kan länka diskontinuerliga flerordsenheter med *I*Link* måste dessa flerordsenheter ändå läggas till i ett efterbehandlingssteg. Så man vinner inte så mycket på att gå via länkfilen jämfört med att extrahera ett lexikon direkt från den supertaggade fdg-filen.

2.6 Efterbehandling

Diskontinuerliga enheter som t.ex. mean – syftar+på i meningen ”Vad **syftar** förkortningen U A **på**?” kan inte länkas med *I*Link*. Därför länkades **mean** mot **syftar** och **på** lämnades

olänkad. Dessa flerordsenheter räknades sedan manuellt och lades till i transferlexikon, svenskt lexikon och superlänkar i efterbehandlingssteget varje gång nya resurser skapades.

Vid efterbehandlingen lades även specifikare nolltaggar in i superlänkarna där något på källspråkssidan var struket. Superlänken N:NOM-SG[]#NULL_LINK byttes ut mot N:NOM-SG#NS0. Dessa nya kategorier lades även in i det svenska lexikonet som möjliga taggar till ”ordet” NULL_LINK.

Ett exempel på specifikare nolltaggar:

Struken enhet	**Nolltagg**
DET:DEF-X[SG] DET:DEF-X[]	DD0
INFMARK[]	INFM0
ING[]	ING0
N:NOM-SG[ATTR]	NS0
N:NOM-SG[DEF&ZERO]	NDS0

Följande skript och resurser användes för efterbehandlingen av resursfilerna:

Skript	Datafiler	Funktion
add-deletions.pl	**deletions**	Lägger in specifikare null-taggar i superlänkarna beroende på vilken tagg som strukits. Mönstrena finns i **deletions**
modify-transfer.pl	**transfer-modifications**	Lägger in MWUer i transferlexikon samt ser till att alla frekvenser är korrekta. MWUerna med frekvens finns i **transfer-modifications**
modify-transfer.pl	**add-entries-to-superlinks**	Lägger in MWUer i superlänkarna samt ser till att alla frekvenser är korrekta. MWUerna med frekvens finns i **add-entries-to-superlinks**
	add-deletions-to-swelex	Diverse nya nolltaggar för NULL_LINK som ska läggas in i svenskt lexikon
	add-entries-to-swelex	Diverse MWUer som ska läggas in i svenskt lexikon.

Figur 9. Efterbehandling av resursfilerna.

3 Utveckling av regler

Följande regeluppsättningar skapades efterhand för att förbättra översättningarna:

- *Supertaggningsregle*r – för engelska och svenska
- *Ordningregler* – transformationsregler
- *Filtreringregler* – filtrerar bort dåliga varianter på målspråkssidan

En förändring av supertaggningsreglerna innebar att alla lexikon måste skapas på nytt. Följande steg upprepades:

1. Supertagga engelska och svenska fdg-filer.
2. Ladda in i *I*Link*. Spara ner nya transferlexikon och superlänkar.
3. Extrahera enspråkiga lexikon för svenska och engelska.
4. Efterbehandling av resursfilerna: diskontinuerliga länkar och null-länkar

Ändringar och tillägg i ordningsregler och filtreringsregler påverkar däremot endast själva översättningsprocessen och inte hur resursfilerna ser ut.

En viktig del i utvecklingen av reglerna var möjligheten att snabbt utvärdera om en ändring inneburit någon förbättring av översättningen. Därför utvärderades översättningen löpande med BLEU-måttet (se avsnittet *Utvärdering*).

Fyra uppsättningar av resursfiler skapades. Första bokstaven anger om resurserna är manuellt generaliserade eller inte (O – original, G- generalized). Den andra bokstaven anger vilken länkning som har använts (O – original (Alignment1), I – improved (Alignment2)).

- *OO* – Resursfiler utvunna ur träningsmaterialet med alignment1.
- *OI* - Resursfiler utvunna ur träningsmaterialet med alignment2.
- *GO* – Resursfiler som i OO med generaliseringar.
- *GI* – Resursfiler som i OI med generaliseringar.

Generaliseringarna för fallen *GO* och *GI* utfördes manuellt och bestod av utökning av alla lexikon. Alla, månader, dagar, siffror, tempus och böjningsformer lades till i lexikon utifrån vad som redan fanns där. Även möjliga taggar för varje ord i de enspråkiga lexikonen lades till. Om en veckodag har två möjliga taggar, lades dessa in för alla veckodagar. Samma sak för siffror, verb, månader, platser etc. Slutligen modifierades transferlexikon och superlänkar för att matcha innehållet i de enspråkiga filerna.

Efter utvärdering av testdata skapades en femte uppsättning resursfiler utökade för att kunna översätta alla meningar i testkorpusen. Den femte uppsättningen bygger på resurserna för *GO* och benämns *GT*.

4 Översättningen

Översättningen med T4F består av följande steg.

	Steg	Resurser
1	Tokenisera engelsk mening	Tokeniseringsskript
2	FDG-tagga mening (Modifiera ordklasstaggningen)	FDG
3	Supertagga meningen	FDG-taggad mening Supertaggningsregler Semantiska kategorier
4	Slå upp varje ord i engelskt lexikon. Hittas inte ordet m. sin supertagg i lexikon lämnas det engelska ordet som det är.	Engelskt lexikon.
5	Hitta översättningsmotsvarigheter till varje ord i meningen.	Svenskt lexikon Transferlexikon Superlänkar
6	Omforma meningen	Orderrules
7	Filtrera bland alternativa ord	Filterrules Syntaktiska kategorier
8	Heuristik	
9	Omforma igen	Orderrules
10	Filtrera igen	Filterrules Syntaktiska kategorier
11	Ge varje möjlig översättning av meningen en sannolikhet och välj det mest sannolika alternativet.	Bigrammodell
12	Sätt ihop sammansättningar (dvs. prefixtag bredvid suffixtag) med bindestreck.	-

Figur 10. Översättningsprocessen.

4.1 Hitta översättningsmotsvarigheter

Efter tokenisering och taggning av källmeningen gäller det att hitta alla möjliga översättningar för orden utifrån den kontext (supertagg) de har. Varje ord slås upp i transferlexikonet och för varje målord slås dess möjliga supertaggar upp i det svenska lexikonet. De målord vars supertaggar förekommer tillsammans med källordets supertagg i superlänkarna behålls som översättningsalternativ.

4.2 Filtrering och transformering

I nästa steg appliceras transformeringsreglerna. Källmeningen ”Do they serve breakfast” måste omformas efter att de rätta ordöversättningarna hittats. Omformningsregeln nedan letar efter mönstret: struket finit verb – vad som helst – verb i presens, och byter plats på huvudverbet och strykningen.

Regel:
IF VFIN0-X-VFR THEN 1:3&3:1

Före transformering	Efter transformering
1 **null_link/VFIN0[]**	1 **serverar/VFR[]**
2 de/PSST[]	2 de/PSST[]
3 **serverar/VFR[]**	3 **null_link/VFIN0[]**
4 frukost/NASI[]	4 frukost/NASI[]

Därefter appliceras filtreringsregler som ska se till att felaktiga översättningar rensas bort, t.ex. översättningar som inte är kongruenta. Här utnyttjas dependensträdet för källspråket eftersom målspråksorden "läggs på" noderna på det engelska dependensträdet. Första regeln nedan betyder t.ex. gå igenom alla översättningar och ta bort obestämda determinerare i neutrum under ett nomen i bestämd form i utrum. De attribut som används är UNDER, ABOVE, 2UNDER, AFTER, BEFORE, LEVELWITH (syskonnoder).

```
RM DIST UNDER NASDN
RM DIST UNDER NASI[N]
RM DIST UNDER NASIX[N]
```

Figur 11. Filtreringsregler.

Före filtrering	Efter filtrering
1 ett/DIST en/DISN	1 en/DISN
2 biljett/NASI[N]	2 biljett/NASI[N]

Figur 12. Filtrering.

Filtreringsregler och transformationsregler fungerar bara då jämförelseordet har fått en slutlig översättning. I fallet nedan kommer inte alternativet ett/DIST att försvinna trots att de båda alternativen ovanför i dependensträdet är sådana att det borde rensas bort.

Före filtrering	Efter filtrering
1 ett/DIST en/DISN	1 ett/DIST en/DISN
2 biljett/NASIX[N] biljett/NASI[N]	2 biljett/NASIX[N] biljett/NASI[N]

Figur 13. Ett filteringsexempel.

För att kunna skriva ett mindre antal, mer generella filtreringsregler skapades överordnade kategorier här kallade Syntaktiska kategorier. T.ex. kategorin PREFIX som innehåller alla förekommande prefixtaggar, NA00PX, AA00PX, PMCPX etc. Dessa kategorier gjorde det också möjligt att formulera ”negativa” filtreringsregler genom att definiera kategorier som NOTPREP som definierar alla taggar utom prepositionstaggarna.

```
NOMEN::NASI NASIG NASIX NAPI NAPIX NAPD NAPDX NASDT …
PREP::S SS ST SAV S2
NOTPREP::NASIG NASI NASIX NAPI NAPIX NAPD NAPDX NASDT …
```

Figur 14. “Syntaktiska kategorier.”

4.3 Bigrammodeller

I det sista filtreringssteget väljs den mening ut som har bäst sannolikhet utifrån en bigrammodell. Tre olika bigrammodeller testades.

Namn	Beskrivning	Exempel
W	ordmodell	vilken sorts flygplan är det
T	taggmodell	WD[N] NASIG[N] NASI[T] VFR[BE] PXST[]
WT	ord+taggmodell	vilken/WD[N] sorts/ NASIG[N] flygplan/NASI[T] är/VFR[BE]

Figur 15. Bigrammodeller.

Sannolikheten för en mening räknades ut enligt nedan:

$$\sum(\log p) * w \qquad (1)$$

där:
p = P(w2 | w1) = frekvens (w1 w2)/ frekvens w1
w = minsta meningslängd/meningens längd

Vikten w är till för att inte kortare meningar ska premieras framför längre. Osedda bigram får sannolikheten 1/10000. I framtida implementeringar kan det vara värt att använda en mer sofistikerad teknik för att approximera dessa sannolikheter, men tillsvidare har den nuvarande lösningen fungerat tillfredsställande.

Exempel.

P(Atlanta | START) = 0,1
P(från | START) = 0,1
P(Atlanta | från) = 0,1
P (till | Atlanta) = 0,1
P(Philadelphia | till) = 0,1

P(från Atlanta till Philadelphia) = (log 0,1 + log 0,1 + log 0,1 + log 0,1) * ¾ = -3
P(Atlanta till Philadelphia) = (log 0,1 + log 0,1 + log 0,1) * 3/3 = -3

4.4 Heuristik

För att ytterligare rensa bland översättningsalternativ används en enkel heuristik för att göra vissa val när endast två översättningsalternativ återstår för ett ord.

1. Om det finns två alternativ där orden är lika medan supertaggarna skiljer sig åt t.ex. flygning/NASI[] flygning/NASIX[] väljs i första hand prefixtagg och sedan suffixtagg (i det här fallet flygning/NASIX[]). Finns ingen prefix eller suffixtagg, väljs ett av orden.

2. Om det finns två alternativ där taggarna är likadana men orden är olika t.ex. flygningar/NAPI[] kvällsflygningar/NAPI[] väljs det ord med högst sannolikhet dvs. det ord som är den vanligaste översättningen av källordet i träningsmaterialet. Sannolikheterna kommer från transferlexikonet. Om orden har lika stor sannolikhet väljs det ena.

Den första heuristiken applicerades bara då bigrammodellen var T medan den andra bara användes då bigrammodellen var W. Ingen heuristik användes för bigrammodellen WT.

5 Utvärdering

BLEU är ett mått för att jämföra en översättning med en referensöversättning (Papineni et al. 2001). Det går i korthet ut på att jämföra n-gram av olika längder mellan en eller flera referensöversättningar och en testöversättning. På det här sättet får man en ganska bra korrelation mellan BLEU och mänskliga bedömare av översättning, både när det gäller korrekthet och flyt. Måttet går från 0 till 1 där 1 betyder att översättningen är perfekt.

5.1 Träningsdata

Resultat på träningsdata – 262 meningar, för alla fyra resursuppsättningar och tre olika bigrammodeller.

	Alignment1		Alignment2	
Ngram-model	Original (OO)	General (GO)	Original (OI)	General (GI)
Word	**0.9782**	0.9724	**0.9826**	0.9756
Tag	0.8937	0.8906	0.8964	0.8908
Word+tag	0.9773	**0.9778**	0.9817	**0.9810**

Figur 16. BLEU för träningsdata.

5.2 Testdata

Resultatet på testdata – 30 meningar. Heuristisk filtrering för alla bigrammodeller.

	Alignment1		Alignment2	
Ngram-model	Original (OO)	General (GO)	Original (OI)	General (GI)
Word	0.7496	0.7715	0.7496	0.7710
Tag	0.6833	0.7030	0.6833	0.7030
Word+tag	**0.7594**	**0.7811**	**0.7594**	**0.7807**

Figur 17. BLEU för testdata m. heuristik för alla bigrammodeller

Sparsamt med heuristik dvs. valet mellan en prefixtagg och en vanlig tagg överlåts åt bigrammodellerna T och WT. Valet mellan ord med likadana taggar lämnas över åt bigrammodellen W.

	Alignment1		Alignment2	
Ngram-model	Original (OO)	General (GO)	Original (OI)	General (GI)
Word	0.7496	0.7715	0.7496	**0.7710**
Tag	0.6718	0.6885	0.6718	0.6885
Word+tag	**0.7507**	**0.7725**	**0.7507**	0.7690

Figur 18. BLEU för testdata m. heuristik för vissa bigrammodeller (W och T)

Bigrammodellen T presterade sämst beroende på att ingen skillnad fanns mellan de svenska taggarna när det gällde sammansättningar som t.ex. flights – flygningar/NAPI förstaklassflygningar/NAPI kvällsflygningar/NAPI. Detta gjorde att kvällsflygningar valdes genomgående i översättningen. WT presterade en aning bättre än W för testkorpusen men skillnaderna var små, antagligen försumbara. För träningskorpusen var WT-modellen endast bättre för de generaliserade testfallen GO och GI.

5.3 T4F vs SCOTS

I jämförelsen nedan presenteras både BLEU-måttet och några av de mått som användes för utvärdering av SCOTS: *precision*, *recall*, *word order* och *identity*. Måttet *identity* anger hur stor andel av översättningarna som var identiska med facit.

System	**Beskrivning**	**Precision**	**Recall**	**Word order**	**Identity**	**BLEU**
OO-wt	Align1, orginalresurser	0.825	0.823	0.846	26,7%	0.759
OI-wt	Align2, orginalresurser.	0.825	0.823	0.846	26,7%	0.759
GO-wt	Align1, generaliserade resurser	0.852	0.852	0.885	30,0%	0.781
GI-wt	Align2, generaliserade resurser	0.847	0.852	0.885	30,0%	0.781
GT-wt	GO + anpassning till testdata	0.924	0.934	0.955	60,0%	0.882
GT-w	GO + anpassning till testdata	0.930	0.938	0.961	60,0%	0.904

Figur 19. Utvärdering för T4F.

System	**Beskrivning**	**Precision**	**Recall**	**Word order**	**Identity**	**BLEU**
SCOTS1-6	Manuell ATIS-taggning	0.802	0.825	0.789	16,7%	0.621
SCOTS2-1	Med sammansättningar	0.802	0.825	0.789	20,0%	0.621
SCOTS2-3	Alla träningsmeningar får en full översättning	0.792	0.817	0.785	16,7%	0.620
SCOTS2-4	Alla träningsmeningar får en acceptabel översättning	0.791	0.811	0.796	16,7%	0.607
SCOTS3-1	Alla ord från test-korpus finns i lexicon.	0.855	0.855	0.822	33,3%	0.651

Figur 20. Utvärdering för SCOTS.

Det testfall som bäst korresponderar med SCOTS2-4 är OO som bygger på samma länkning och endast innehåller ord och konstruktioner från träningsdata.

6 Diskussion

6.1 Effektivitet

Ett potentiellt problem just nu är att vi för att hitta den mest sannolika översättningen först genererar alla möjliga permutationer av de återstående översättningsalternativen och därefter beräknar dess sannolikhet utifrån bigramfrekvenserna. För andra domäner med längre meningar t.ex. Access kan detta bli ohållbart, här är det bättre att använda effektivare algoritmer som Viterbi eller A*-sökning.

6.2 Utvärdering

BLEU-måttet är behäftat med vissa problem. Meningar som är kortare än det längsta n-gram som används för jämförelse, får 0 i omdöme. I den här utvärderingen gällde detta meningar kortare än 4 ord t.ex. Tidiga flygningar.

6.3 Felkällor

En av slutsatserna efter SCOTS var att testmaterialet inte täckte träningsmaterialet särskilt bra. Många nya ord och översättningar introducerades, vissa följde naturligt ur träningsmaterialet och fanns därför i de generaliserade resurserna (GO och GI) t.ex. alla veckodagar och månader, medan andra var helt nya.

När det gäller FDG-parsningen finns det inte så stora möjligheter att påverka hur resultatet blir. Ett sätt är som nämnts tidigare att undertrycka möjliga tolkningar för ord i ett anpassat lexikon vilket i sin tur skulle kunna innebära nya problem om man inte är försiktig. Möjligheten att anpassa sitt lexikon finns dock bara för den engelska parsern.

Filtreringsreglerna fungerar endast om det bara finns ett alternativ kvar på omgivande ord. Filtreringen skulle kunna göras kraftfullare så att t.ex. en determinerare i singular filtreras bort då *alla* orden ovanför i dependensträdet är nomen i plural.

Liksom i SCOTS gör kravet på 1:N länkning att ”evening” i ”evening flights” stryks och ”flights” länkas till ”kvällsflygningar”, en översättning som inte är optimal. Poängen här var dock att använda samma länkning som tidigare för att kunna göra en rättvis jämförelse mellan systemen.

6.4 Utvecklingsmetod

Många av stegen vid tillverkningen av resursfilerna skulle försvinna om de funktioner som behövs byggs in i *I*Link*. Detta inkluderar funktionalitet för att göra diskontinuerliga länkar och specificera noll-taggar.

När det gäller utvecklingen av systemet dvs. att bygga de regler som behövs finns också en hel del att förbättra. Till detta skulle man vilja ha ett verktyg som ger direkt feedback på de ändringar man gör. Om en ny supertaggningsregel läggs till bör man kunna testa dess effekt inte bara med BLEU-måttet (som ju tar de färdiga översättningarna som gått igenom det sista statistiska steget som sina testdata) utan med ”interna” mått t.ex. mått på ambiguiteter i översättningen av träningsdata, eller mått på ambiguiteten hos lexikon.

6.5 Byte av domän

De flesta supertaggningsregler och filtreringsregler kan troligen flyttas över till en ny domän. De särdrag i omgivningen som är betydelsefulla för ordens översättning bör gälla även över domängränserna. På samma sätt bör filtreringsreglerna överensstämma eftersom de mestadels bygger på vilka kombinationer som är grammatiskt riktiga/felaktiga i svenska. Formuleringen av reglerna måste självklart anpassas till den tagguppsättning som används.

7 Referenser

Ahrenberg, L. (2000). “Superlinks: A new approach to constraining transfer in machine translation”, Linköping University, PLUG Project Report

Jonsson, H. (2001). ”Exploring Superlink Constrained Lexicalist Transfer for Empirical Machine Translation”, Master thesis, Linköpings universitet.

Papaneni, K. et al. (2001). ”BLEU: a Method for Automatic Evaluation of Machine Translation”, IBM Research Report, RC22176.

Tapanainen, P. and Järvinen, T. (1997). "A non-projective dependency parser", presented at ANLP-97 ACL, Washington, D.C.